Awareness as Relevance Selection: A Causal Framework for Attention, Internal Feedback, and Artificial Intelligence

Denis Saklakov
Robotech Frontier Hub
ds@robotechfrontierhub.com
ORCID: 0009-0005-0756-7303

April 2026

Abstract

The strong form of this framework is falsified if selective intervention on a decoded relevance variable produces no disproportionate impairment in switching cost or calibration error relative to matched perturbations of attention, metacognition, and generic latent state. Everything that follows is structured around that test. Biological intelligence does not process all available data equally. It continually selects a restricted subset of signals as consequential for control, inference, and action. This paper argues that awareness can be studied as a causal regime of relevance selection rather than as a purely philosophical category or a loose synonym for attention. In the proposed framework, attention is a gating mechanism that regulates access to computation, relevance is a criterion-dependent estimate of expected control value, internal feedback is the recurrent return of compressed relevance state into the system's latent dynamics, metacognition is a partially overlapping but separable family of self-evaluative variables, and functional control transitions occur when recurrent relevance feedback changes subsequent processing in a causally testable manner. We propose and test the hypothesis that a relevance feedback regime emerges when a system jointly performs three operations: it estimates relevance under an explicit criterion, compresses the selected state into a persistent internal variable, and re-injects that variable to modulate confidence, switching, robustness, and policy stability. This formulation is designed to separate attention from awareness, awareness from reportability, and relevance from reward prediction error, predictive coding, or global broadcast alone. The paper develops a formal causal architecture, formulates falsifiable hypotheses, defines intervention points, and proposes biological, artificial, and cross-domain experiments. It also argues that biological and artificial systems should differ systematically because their relevance criteria differ: biological cognition is strongly shaped by survival, homeostasis, threat, social salience, and temporal urgency, whereas artificial systems can be optimized for uncertainty reduction, anomaly detection, long-horizon coherence, external goals, or hybrid criteria. The proposed program aims to establish a comparative science of functional control grounded in measurable latent variables rather than metaphor. The strong form is falsified if (i) no candidate relevance subspace predicts held-out control above difficulty, arousal, and attention controls across two independent task paradigms, or (ii) $\mathrm{do}(r_t)$ produces no selectively disproportionate impairment relative to matched $\mathrm{do}(a_t)$, $\mathrm{do}(m_t)$, and $\mathrm{do}(h_t)$ interventions. The framework does not retreat to awareness-like language as a hedge - it earns the term awareness by defining it as the causal regime that survives the four-arm intervention test.

Keywords: awareness, attention, relevance selection, causal inference, metacognition, predictive

processing, reward prediction error, global neuronal workspace, artificial intelligence, cognitive neuroscience

Executive Summary

The Single Fastest Test: For a skeptic with existing equipment, take any attention task with volatile cue switching. Add distractors. Record from prefrontal cortex. Decode a low-dimensional population axis that predicts switching on held-out trials. Artificially perturb that axis, using optogenetics in rodents or noise injection or clamping in AI, during the post-cue window. If switching cost or calibration error increases disproportionately relative to basic sensory discrimination, and more so than under matched do(a_t), do(m_t), or do(h_t), the framework satisfies the falsification conditions in Sections 4.6 and 6.4. If not, see falsification conditions in Sections 4.6 and 6.4.

The decisive empirical test of this framework is the four-arm causal perturbation study in Section 6.4. If selective intervention on r_t produces a selectively disproportionate impairment in switching cost, calibration error, or distractor resistance relative to matched perturbations of a_t, m_t, and nonspecific latent state h_t, the theory satisfies the falsification conditions in Sections 4.6 and 6.4. If the ratio of switching or calibration deficit to basic sensory discrimination deficit under do(r_t) is indistinguishable from the same ratio under matched do(a_t) or do(h_t) interventions, the theory fails in its strong form.

The two highest-risk and highest-payoff hypotheses are H7 and H8. If criterion identity is encoded in relevance geometry, then awareness-like control becomes comparable across systems in a principled way. If compression obeys an inverted-U relation, then the theory predicts a concrete architectural law for when relevance feedback becomes stable, interpretable, and causally useful in both biological and artificial systems.

1 Introduction

1.1 Problem statement

Biological intelligence does not process all available data equally. It selects what matters, and that selection is constrained by inherited and learned pressures tied to survival, control, and uncertainty reduction. Classical work on attention established that neural systems allocate resources nonuniformly across stimuli, locations, tasks, and timescales, and later work refined this into interacting orienting, alerting, and executive networks [1, 2, 3]. Yet selective routing alone does not explain why some signals become internally decisive, why some selected states alter confidence or switching, or why perturbing some latent variables changes policy stability more than perturbing attention alone.

A note on the origin of the hypothesis. The distinction between attentional gating and recurrent relevance feedback that structures this framework did not arise purely from the theoretical literature. It was first encountered as a phenomenological observation during intensive Vipassana meditation practice, in which sustained body scanning produces a qualitatively distinct control state — one that persists across time, reorganizes the salience of incoming signals, and demonstrably alters subsequent behavioral policy in ways that moment-to-moment attention alone does not predict. This observation is not offered as evidence. It is offered as the empirical intuition that preceded the formalism and motivated the specific architectural separation of attention, relevance, and internal feedback that the framework proposes. The value of that

intuition, like any intuition in science, is entirely conditional on whether the formal predictions it generated survive causal testing. The present paper is that test.

1.2 Core idea

Awareness is treated as a recurrent control regime in which selected information is returned to the system in a compact internal form that changes what the system does next. Attention is part of the mechanism, but not the whole mechanism. Attention gates candidate content. Relevance estimates which content has expected control value under a criterion. Internal feedback returns that selected state into the latent dynamics. Functional control appears when the returned variable alters subsequent processing in a way that is decodable, manipulable, and behaviorally specific.

This means awareness, in the present framework, is not a binary property that a system either has or lacks. It is a family of criterion-dependent recurrent control regimes, graded, measurable, and comparable across biological and artificial systems.

1.3 Scientific gap

Current artificial intelligence work often studies attention as a routing mechanism or performance feature, but not as part of a measurable relevance-selection system. In cognitive neuroscience, attention, confidence, metacognition, predictive processing, reward prediction error, and conscious access are often studied in adjacent but only partially integrated literatures [4, 5, 6, 7, 10, 12]. In machine learning, attention distributions are often over-interpreted even though influential work has shown that standard attention weights are neither sufficient nor reliably causal explanations of model behavior [16, 17, 18]. What remains unspecified is a causal account that distinguishes attentional gating, relevance estimation, internal feedback, metacognitive monitoring, and relevance-feedback state transitions as separable components of one causal architecture.

1.4 Research question

Can internal relevance-selection mechanisms be identified, measured, and causally tested in neural systems, and how do biologically inherited versus engineered relevance criteria change functional control behavior?

1.5 Contribution

This paper proposes a formal and experimentally tractable framework in which attention selects candidate signals, a relevance function estimates their control value under a criterion, internal feedback returns a compact relevance state, metacognitive variables partially read out that state, and criterion-dependent recurrent dynamics determine what kind of relevance feedback regime emerges. The contribution is not a claim that biological and artificial awareness are identical. It is a proposal for comparing them under a common causal language and a common perturbation logic.

A useful retrodictive anchor comes from masked-perception and masked-priming dissociations. Subjects can show above-chance forced-choice discrimination or action priming from a masked stimulus while reporting little or no confidence in having seen it [8, 21, 22]. Attention-as-routing accounts that do not distinguish gating from recurrent control predict closer covariation between selective processing and report than the present framework requires, whereas higher-order

accounts often treat reportability as the decisive variable. The present framework predicts a third pattern: relevance feedback can still be present strongly enough to support above-chance control while metacognitive readout is degraded, because r_t and m_t are separable components of the architecture. This retrodiction follows directly from the structural prediction that $\mathrm{do}(m_t)$ and $\mathrm{do}(r_t)$ should yield different behavioral phenotypes. The masked-priming and relative blindsight findings are therefore retrodictions of the framework, not just motivations for it.

Awareness is not a synonym for attention, reportability, or consciousness in this framework. It names the causal regime - criterion-sensitive, recurrently injected, selectively perturbable - that this paper defines, measures, and distinguishes from its neighbors.

2 Conceptual Background

2.1 Survival-shaped relevance in biology

Biological cognition evolved under the joint pressures of energy limitation, bodily vulnerability, temporal uncertainty, and action selection. As a result, nervous systems inherit strong priors about what is worth noticing: threat, nourishment, novelty, motion, temperature, pain, social cues, temporal contingency, and signals predictive of control-relevant outcomes. This does not mean that all salience reduces to reward or all awareness reduces to utility maximization. It means that biological relevance is historically shaped by control consequences for the organism. Dopaminergic reward prediction error is one special case of this larger adaptive logic, not the whole of it [12, 14, 15]. Threat, conflict, surprise, bodily deviation, and strategic uncertainty can dominate control before any scalar reward outcome is resolved.

2.2 Attention as a selection mechanism

Attention does not create awareness by itself. It filters available data so that some signals become candidates for deeper processing, memory updating, and action. Classical and contemporary work supports a multi-component view involving orienting, executive control, working memory interaction, top-down sensitivity control, and bottom-up filtering [1, 2, 3]. A mechanistic understanding of attention is therefore necessary for a theory of awareness-like control, but still insufficient. Routing can occur without robust introspective access, and global access can fail even when some sensory or decision-related processing continues. Therefore a theory that equates attention with awareness collapses gating into control and cannot explain dissociations among selective processing, confidence, switching, and reportability.

2.3 Awareness as measurable feedback

Awareness is treated here as the recurrent return of testable internal signals, analogous in formal role, not in phenomenology, to bodily feedback channels such as pressure, temperature, or positional deviation. The point is operational: if an internal variable is awareness-relevant, it must be decodable from system state, it must covary with specific control properties, and direct perturbation of that variable must alter downstream behavior in a characteristic way. This sharply separates the present proposal from philosophical uses of the term awareness that lack intervention criteria. Under this framework, awareness-like claims are scientific only if they survive causal testing.

2.4 Relevance functions in artificial systems

Unlike biology, artificial systems need not inherit survival bias. They can be trained with relevance criteria such as anomaly detection, coherence maintenance, long-horizon planning, active uncertainty minimization, multimodal consistency, or externally specified reward. This engineered flexibility is scientifically valuable because it allows relevance criteria to be manipulated directly. If awareness-like control depends on criterion-sensitive relevance feedback, then one should be able to change the internal geometry, decodability, and behavioral effects of relevance variables by changing the criterion while holding architecture approximately fixed. This is one of the main reasons artificial systems are useful for theory testing rather than analogy.

The reinforcement learning literature on intrinsic motivation has already operationalized criterion-dependent relevance in artificial systems through curiosity-driven exploration, empowerment, and information-gain objectives [29, 30, 31, 32]. The key distinction is that intrinsic motivation research primarily targets exploration bonuses and learning efficiency, whereas the present framework targets the decodability, compressibility, and causal perturbability of a persistent relevance state as a control variable. The gridworld experiment and bottleneck sweep are designed to test properties that intrinsic-motivation frameworks do not typically measure, specifically the disproportionate perturbation phenotype and criterion-geometry reorganization.

2.5 Operational definitions

The manuscript uses the following terms operationally rather than rhetorically.

Henceforth, candidate r_t-like subspace refers to any low-dimensional latent variable that jointly satisfies the four identification criteria in Section 4.4: decodability, temporal persistence, criterion sensitivity, and causal efficacy. This is an operational construct, not a claim that such a subspace has been isolated in any existing system. No claim is made that a_t, r_t, or m_t correspond to anatomically or architecturally isolated modules; they are distinguished by their intervention profiles in Section 4.5.

3 Research Goals and Hypotheses

3.1 Central question

The central question is whether relevance-selection functions can be defined and manipulated in neural systems so that awareness-like feedback becomes observable, decodable, and causally testable.

3.2 Hypothesis set

H1: Decodable latent relevance. A system's internal relevance signal can be decoded from hidden or neural population states more robustly than raw attention maps alone. The mechanism is that relevance is a compressed recurrent state variable rather than a momentary routing pattern. The predicted observation is that decoded relevance forecasts later switching, calibration, distractor resistance, and confidence better than attention weights. The main rival explanation is that decoded relevance is merely a proxy for task difficulty, arousal, or global activation.

H2: Causal specificity of relevance perturbation. Directly perturbing the relevance signal will change behavior in specific and measurable ways that differ from perturbing attention gates or generic network activity. The mechanism is that relevance feedback enters latent control dynamics downstream of selection. The predicted observation is selective impairment in switching thresholds, confidence calibration, context maintenance, or distractor resistance without uniform collapse of all performance. The main rival explanation is that any latent perturbation could produce these effects nonspecifically.

H3: Criterion dependence of awareness-like control (Criterion-sensitive reorganization). Different relevance criteria produce different patterns of awareness-like control. The mechanism is that the criterion changes what the system treats as expected control value. The predicted observation is criterion-dependent reorganization of latent state geometry, calibration, vulnerability, and failure mode. The main rival explanation is that differences merely reflect retraining noise or distribution shift.

H4: Dissociation between relevance and explicit confidence. Relevance and metacognitive confidence overlap but do not coincide. The mechanism is that confidence is one readout of internal state quality, whereas relevance is a broader control variable that can encode urgency, conflict, anomaly, or homeostatic consequence even in the absence of explicit self-report. The predicted observation is that some perturbations alter switching and robustness without linearly altering reported confidence. The main rival explanation is measurement noise in confidence reports.

H5: Biological salience exceeds reward prediction error. In biological systems, survival-linked relevance should dominate latent control beyond reward prediction error alone. The mechanism is that biological relevance includes bodily threat, urgency, and social salience not reducible to scalar reward discrepancy. The predicted observation is that decoded relevance in neural recordings tracks bodily deviation, conflict, and socially salient signals after controlling for expected reward. A concrete discriminating test holds expected reward constant while varying hazard rate or interoceptive load parametrically. If decoded relevance tracks the manipulation after residualizing out estimated reward prediction error, H5 survives. If it does not, H5 fails while the broader framework remains intact. The main rival explanation is hidden differences in subjective value.

H6: Recurrent relevance outperforms wider attention. Adding recurrent relevance feedback should improve calibration, strategic switching, and distraction resistance more than simply widening attention. The mechanism is that persistent relevance state supports control stability in a way static routing cannot. The predicted observation is that architectures with explicit relevance state outperform parameter-matched wider-attention baselines on volatile, distractor-rich, or delayed-report tasks. The main rival explanation is parameter count, regularization, or optimization differences.

H7: Criterion identity is encoded in relevance geometry (Criterion decodability from relevance geometry). If relevance is genuinely criterion-bound, then the geometry of the decoded relevance manifold should classify which criterion the agent is currently optimizing even when stimuli and base architecture are matched. The predicted observation is above-chance

decoding of criterion identity from relevance subspace but not from raw attention weights alone. The rival explanation is that auxiliary context or task labels leak criterion identity.

H8: Compression is necessary for awareness-like control. A relevance signal that is not compressed into a relatively low-dimensional bottleneck will be less stable and less causally interpretable. The core claim is that awareness-like control depends on a compressed recurrent state rather than a diffuse echo of the full latent manifold. The compression requirement is not a heuristic preference — it follows as a theoretical necessity from the architecture itself. The bottleneck width is bounded below by the minimum number of bits required to distinguish criterion-relevant signal classes from criterion-irrelevant ones: a bottleneck narrower than this threshold cannot preserve the distinctions that make relevance feedback causally meaningful. It is bounded above by the point at which r_t becomes entangled with h_t, losing the intervention-specific separability that the four-arm perturbation study requires. The inverted-U performance curve predicted by H8 is therefore not an empirical guess — it is the expected shape of any compression function operating between these two bounds. The sweep over $\dim(r_t)$ tests where the bounds fall empirically. The shape of the curve is the scientific claim; the location of the peak is the empirical finding. This compression logic is grounded in the information bottleneck principle in representation learning, where optimal representations maximally compress input while maximally predicting task-relevant output [23, 24]. The test is straightforward: models M2 and M3 from Table 2 should be trained with systematically varied bottleneck widths, for example sweeping $\dim(r_t)$ from 1 to $\dim(h_t)/2$, and evaluated on calibration and switching cost to recover the predicted inverted-U. The main rival explanation is that any such curve reflects generic architectural regularization rather than compression-specific control.

H9: Criterion-invariant structure (Criterion-invariant subspace). Across criterion changes, a subset of internal representations z^* retains decodability, predictive power for control outcomes, and causal efficacy. Formally, for candidate variable z^*:

$$\forall c_i, c_j : \ \mathrm{Role}(z^* \mid c_i) \approx \mathrm{Role}(z^* \mid c_j), \tag{1}$$

where Role includes prediction of control outcomes and effect under intervention, measured by a pre-registered decoding accuracy ratio or RSA correlation coefficient, with invariance defined as falling within a tolerance band across all criterion pairs tested.

The criterion-invariant subspace z^* carries a specific geometric signature that distinguishes it from shared variance or statistical artifact. Its geometry is predicted to be more stable under criterion change than its predictive power for control outcomes. That is, the manifold structure of z^* should persist even when criterion shifts reorganize the criterion-sensitive subspace identified by H3 and H7, and this persistence should be measurable as RSA correlation above a pre-registered tolerance band across all criterion pairs tested. The speculative implication — that the deepest layers of z^* may reflect mathematical or physical necessity rather than evolutionary contingency — is downstream of this empirical prediction, not prior to it. If z^* satisfies the stability criterion across a sufficiently broad criterion distribution in both biological and artificial systems, the question of whether its structure reflects universal constraints on causal information processing becomes a legitimate theoretical target for subsequent work. The present paper does not answer that question. It establishes the empirical conditions under which it becomes worth asking.

Invariance is defined operationally as a pre-registered tolerance band, for example RSA

correlation > 0.7 across all criterion pairs. Failure condition: see falsification conditions in Sections 4.6 and 6.4.

The hypothesis does not treat z^* as a philosophical absolute. It treats it as the expected residue of optimization under a broad distribution of criteria: representations that are control-relevant under any criterion in the system's operating range because the conditions that shaped the system, physical, energetic, temporal, social, imposed common structure across all of them. Biological systems are therefore predicted to exhibit a richer and more stable z^* than artificial systems optimized under narrow criterion families. Artificial systems trained on broad, diverse criterion distributions are predicted to converge toward biological z^* structure.

H9 is complementary to, not contradictory with, H3 and H7. H3 and H7 concern the criterion-sensitive subspace that reorganizes when c changes. H9 concerns the criterion-invariant subspace that persists through that reorganization. Both subspaces are predicted to coexist within the same latent manifold.

For the binding standard, see falsification conditions in Sections 4.6 and 6.4. The experimental cost of testing H9 is near zero, as it is a second analysis on data already collected for H3 and H7.

3.3 Biological-artificial comparison

The asymmetry between biological and artificial relevance criteria is not merely a background assumption — it generates a specific structural prediction about relevance manifold geometry. Biological relevance criteria were shaped by selection pressures that imposed asymmetric weighting on certain signal classes: threat, homeostatic deviation, and social salience were catastrophically consequential at base rates far lower than their representation in biological relevance manifolds would suggest. This asymmetry means that the geometry of biological z^* is predicted to be non-isotropic in a specific way — threat and homeostatic signals should occupy disproportionately large and stable manifold regions relative to their base-rate frequency in any given task environment. An artificial system trained on reward-matched criteria without survival-asymmetric weighting should exhibit a comparatively flatter, more isotropic relevance geometry under RSA. This is a directly testable cross-domain prediction: RSA comparison of biological and artificial relevance manifolds under matched task statistics should reveal the predicted geometric asymmetry. If biological and artificial manifolds are statistically indistinguishable in geometry under matched reward, the survival-shaping hypothesis fails while the broader relevance-feedback architecture remains intact.

3.4 Approximation strategies for counterfactual relevance

The non-computability of Equation 5 in closed form has been noted as a limitation. It should instead be stated as a comparative advantage. Every major competing framework relies on quantities that are equally or more non-computable in practice: integrated information Φ requires exponential combinatorial search over system partitions; expected free energy under active inference requires a generative model whose parameters are never fully specified in biological implementations; global ignition threshold has no agreed computational definition outside of simplified network models. Equation 5 is uniquely advantaged because it admits a tractable approximation cascade with explicit causal intervention targets, and because in artificial systems the counterfactual is not approximate at all — it is computed exactly by clamping r_t and evaluating the resulting trajectory. The approximation gap for $\mathrm{Rel}_{\mathrm{fb}}$ is smaller than the approximation gap for any rival quantity in at least one implementation domain. That is not

a limitation dressed up as a virtue. It is a genuine methodological advantage that competing theories do not share.

The approximation cascade below operationalizes this advantage concretely. In biological systems, the exact interventional quantity cannot be computed directly in every case; in artificial systems, it can be approximated or directly evaluated by clamping the relevant variable. Practical estimation therefore proceeds through an approximation cascade. In artificial systems, Monte Carlo rollout can estimate the expected future-control difference between trajectories in which a candidate signal is selected and recurrently returned and matched trajectories in which the signal is clamped, masked, or ablated. When full rollout is too expensive, local influence functions can estimate the sensitivity of switching, calibration, distractor susceptibility, or policy stability to small perturbations of candidate latent variables. In both biological and artificial systems, targeted ablation or stimulation can supply a causal lower-bound estimate by testing whether perturbing the decoded candidate variable changes held-out control outcomes in the predicted direction. These procedures do not make Equation 5 exact; they specify the tractable bridge from the formal definition to the intervention program used in Sections 5 and 6.

No existing awareness theory provides a task-local counterfactual control-gain quantity for a decodable, selectively perturbable latent variable. Active inference specifies expected free energy as a computable quantity; the present framework specifies something narrower: a criterion-bound control-gain estimate that must be decodable as a low-dimensional variable, geometrically reorganized by criterion change, and selectively interruptible with a disproportionate behavioral phenotype.

4 Formal Causal Architecture

4.1 State variables

Let x_t denote exteroceptive input, i_t interoceptive or internal bodily state, h_t the latent system state, a_t an attentional gating vector, r_t a latent relevance state, m_t a metacognitive state, c the relevance criterion, and u_t the policy or action output. Let $\tau_{t:t+T}$ denote a future trajectory over a finite horizon T.

The functions ϕ, A, Ψ, M, and G are schematic. Their specific forms are underdetermined by the framework and must be instantiated and pre-registered within each experimental implementation. The framework's falsifiability rests on intervention logic in Section 4.5, not on any particular functional form.

The underspecification is deliberate: it forces each experimental implementation to pre-register its functional forms before data collection, making the falsification conditions binding rather than retroactive. The underspecification allows different implementations, linear, nonlinear, recurrent, or convolutional, to be tested against the same causal logic. Pre-registration of functional forms before data collection is the safeguard against post-hoc fitting.

4.2 Selection function

The architecture begins with candidate state construction:

$$z_t = \phi(x_t, i_t, h_{t-1}). \tag{2}$$

Attention gates candidate components:

$$a_t = A(z_t, h_{t-1}, c), \tag{3}$$

where a_t is a routing distribution or gating tensor over candidate features. The gated candidate state is

$$\tilde{z}_t = a_t \odot z_t. \tag{4}$$

To define relevance rigorously, the manuscript replaces the vague notion of "importance" with a counterfactual control-gain quantity. Let $\mathcal{L}_c(\tau_{t:t+T})$ be the criterion-defined cumulative loss over future behavior. Then for a candidate signal $s \subseteq \tilde{z}_t$,

$$\mathrm{Rel}_t(s; c) = \mathbb{E}\left[\mathcal{L}_c(\tau^{\neg s}_{t:t+T}) - \mathcal{L}_c(\tau^{s}_{t:t+T}) \mid h_{t-1}, c\right]. \tag{5}$$

Practical approximation strategies for this quantity, including Monte Carlo rollout, local influence functions, and targeted ablation, are specified in Section 3.4. Here $\tau^{s}_{t:t+T}$ denotes the future trajectory if s is selected and recurrently returned to the system, and $\tau^{\neg s}_{t:t+T}$ denotes the matched counterfactual trajectory when its contribution is clamped or removed. Relevance is therefore expected *improvement in future control under criterion* c, not mere salience, surprise, or activation magnitude.

To make the framework's central distinction precise, define two component quantities. Let $\mathrm{Rel}_{\mathrm{sel}}(s; c)$ denote the expected control gain from selecting s alone, without recurrent return. Let $\mathrm{Rel}_{\mathrm{fb}}(s; c)$ denote the incremental expected control gain from recurrently returning s to latent dynamics, conditional on selection. Then:

$$\mathrm{Rel}(s; c) = \mathrm{Rel}_{\mathrm{sel}}(s; c) + \mathrm{Rel}_{\mathrm{fb}}(s; c). \tag{6}$$

The present framework identifies relevance with $\mathrm{Rel}_{\mathrm{fb}}$, not $\mathrm{Rel}_{\mathrm{sel}}$. Selection is attention's contribution. Recurrent return is awareness's contribution. This separation is the operational core of the paper and the basis on which attention and awareness-like control are distinguished throughout.

A compressed relevance state is then formed:

$$r_t = \Psi(\{\mathrm{Rel}_t(s; c)\}_{s \in \mathcal{S}_t}, h_{t-1}, m_{t-1}), \tag{7}$$

where $\mathcal{S}_t$ is the set of candidate signals and $\dim(r_t) \ll \dim(h_t)$.

One principled way to express the compression requirement is as an information-bottleneck objective [23, 25]:

$$q^\star(r_t \mid \tilde{z}_t, h_{t-1}, c) = \arg\min_q I_q(R_t; \tilde{Z}_t, H_{t-1}) - \beta\, I_q(R_t; Y^{\mathrm{ctrl}}_{t:t+T}, C), \tag{8}$$

where $Y^{\mathrm{ctrl}}_{t:t+T}$ denotes future control-relevant outcomes such as switching, calibration, or policy stability. Equation 8 states the intended mechanism of compression: r_t should retain as much information as possible about future control under the criterion while discarding irrelevant latent detail.

4.3 Feedback loop

Internal feedback is the recurrent re-entry of relevance into latent dynamics:

$$m_t = M(\tilde{z}_t, r_t, h_{t-1}), \tag{9}$$

$$h_t = G(\tilde{z}_t, r_t, m_t, h_{t-1}), \tag{10}$$

$$u_t \sim \pi(\cdot \mid h_t, r_t, m_t). \tag{11}$$

In this architecture, attention and relevance are not identical. Attention determines access. Relevance determines control priority. Metacognition estimates aspects of internal state quality, such as confidence, uncertainty, or expected error. Awareness-like state transitions occur when recurrent injection of r_t changes future latent dynamics and policy in a way that is proportionally distinguishable from matched do(a_t), do(m_t), and do(h_t) interventions. A recurrent variable that produces only diffuse or proportionally uniform impairment does not meet this standard.

The recurrent injection of r_t into h_t implies a prediction about temporal persistence. If r_t is genuinely control-relevant rather than a momentary routing signal, its autocorrelation structure should differ systematically from that of a_t. Specifically, the framework predicts that the autocorrelation of r_t decays more slowly than that of a_t across task-relevant timescales, and that the mutual information between r_t at time t and policy output u_{t+k} remains above chance for at least as many timesteps as the criterion horizon τ. In tasks with trial structure, this predicts that decoded relevance from the post-cue epoch should predict behavioral outcomes in the response epoch better than decoded attention weights do, and that this predictive advantage should grow with task volatility, because volatile environments require longer relevance horizon to stabilize switching decisions. In tasks with sustained context, such as the long-horizon consistency task in Section 5.3, the relevance state should remain measurably above the noise floor across the full inter-reversal interval.

4.4 Identification logic

Because a_t, r_t, and m_t may occupy partially overlapping subspaces of a shared latent manifold rather than cleanly separated modules, the framework does not require architectural isolability. It requires experimentally useful dissociation, operationalized through the following four criteria that a candidate relevance dimension must jointly satisfy:

1. **Decodability:** a low-dimensional summary of r_t can be extracted from latent or neural population activity above matched null and difficulty controls.
2. **Temporal persistence:** r_t predicts future control-relevant state better than instantaneous routing measures.
3. **Criterion sensitivity:** its geometry reorganizes when c changes while sensory input statistics are held as constant as possible.
4. **Causal efficacy:** perturbing r_t changes later behavior more specifically than matched perturbations of routing, confidence readout, or generic latent activity.

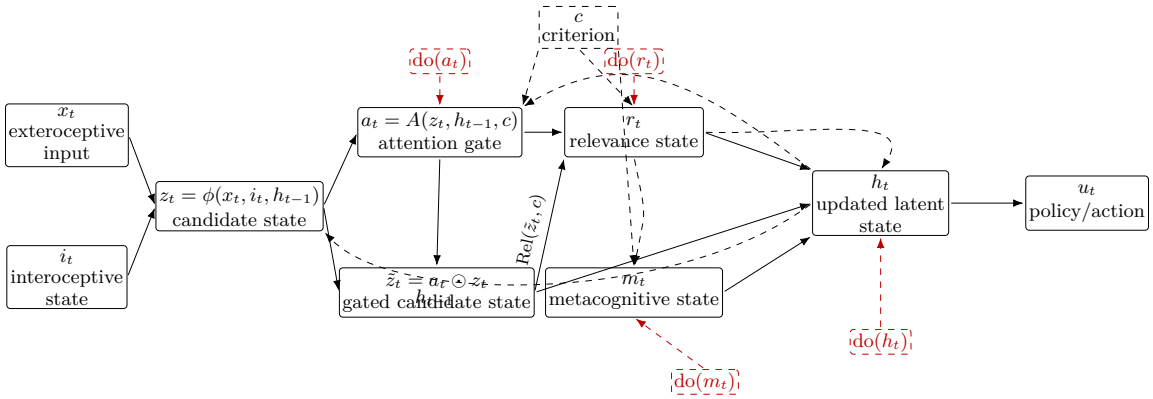

Figure 1: Causal architecture of the relevance-selection framework. Solid arrows denote forward information flow. Dashed arrows denote recurrent feedback. Red intervention markers indicate the four perturbation points used in the causal perturbation study. The criterion c enters the system at attentional gating, relevance estimation, and metacognitive monitoring. This diagram is not an illustrative metaphor and does not represent anatomical or computational modularity. It fixes the intervention logic of the paper. Any implementation that cannot instantiate the four perturbation points $\mathrm{do}(a_t)$, $\mathrm{do}(r_t)$, $\mathrm{do}(m_t)$, and $\mathrm{do}(h_t)$ as experimentally distinguishable operations does not constitute a test of the framework.

The same sensory input can pass through attention without becoming a persistent relevance state, and a relevance estimate can alter latent dynamics without being identical to metacognitive confidence.

The four identification criteria do not guarantee uniqueness. Multiple latent variables could in principle satisfy decodability, temporal persistence, criterion sensitivity, and causal efficacy simultaneously. The framework identifies candidate relevance dimensions subject to further causal discrimination, not a provably unique decomposition.

Non-uniqueness is a real identifiability constraint. The four criteria do not guarantee a unique decomposition. They define an equivalence class: any latent variable satisfying decodability, persistence, criterion sensitivity, and causal efficacy is a candidate r_t. If multiple candidates exist and are empirically indistinguishable under current methods, that defines the limit of first-generation testing and the agenda for the next generation. Uniqueness is not required for the causal perturbation program to proceed.

4.5 Causal intervention points

The framework yields at least four distinct intervention points:

$$\mathrm{do}(a_t \leftarrow a'_t), \tag{12}$$

$$\mathrm{do}(r_t \leftarrow r'_t), \tag{13}$$

$$\mathrm{do}(m_t \leftarrow m'_t), \tag{14}$$

$$\mathrm{do}(h_t \leftarrow h'_t). \tag{15}$$

$\mathrm{do}(r_t)$ is a hypothetical target in the formal model. In biological implementations, it is approximated via closed-loop optogenetic suppression of a candidate neural subspace as described in Section 6.1. In artificial implementations, it is exact via clamping or perturbation of the r_t variable as described in Section 6.2.

The evidentiary hierarchy is as follows. In artificial systems, $\mathrm{do}(r_t)$ is exact - the variable

is directly clamped or perturbed. In rodent systems, $\mathrm{do}(r_t)$ is approximated via closed-loop optogenetic suppression of a decoded candidate subspace. In human systems, $\mathrm{do}(r_t)$ is further approximated via state-triggered TMS to a broad neural region. Claims scale accordingly: exact causal demonstration in AI, strong causal approximation in rodents, convergent suggestive evidence in humans.

A strong claim for awareness-like relevance requires that interventions on r_t produce specific effects not reducible to changing a_t or broadly damaging h_t. Formally, for an outcome Y such as switching cost or calibration error,

$$\Delta_Y^{(r)} = \mathbb{E}[Y_{t:t+T} \mid \mathrm{do}(r_t = r_t + \delta)] - \mathbb{E}[Y_{t:t+T} \mid \mathrm{do}(r_t = r_t)] \tag{16}$$

must be both non-zero and phenotypically distinct from matched $\Delta_Y^{(a)}$, $\Delta_Y^{(m)}$, and nonspecific latent perturbation.

4.6 Observable outputs and falsifiability

Observable outputs include accuracy, switching cost, reaction time, calibration, distractor vulnerability, confidence readout, latent manifold structure, and intervention sensitivity profiles. The distinctive signature of awareness-like relevance is a joint pattern: decodable low-dimensional relevance state, temporal persistence, criterion dependence, and causal influence on control variables beyond routing alone.

The strong form of the theory is falsified if no candidate relevance subspace predicts held-out control outcomes above difficulty, arousal, value, and attention controls, and this failure persists across at least two independent task paradigms with sufficient statistical power; if $\mathrm{do}(r_t)$ produces no selectively disproportionate impairment relative to matched $\mathrm{do}(a_t)$, $\mathrm{do}(m_t)$, and $\mathrm{do}(h_t)$ interventions; or if changing the criterion fails to reorganize the relevance manifold while leaving architecture constant.

4.7 Predictions Checkable Without New Data

The framework generates predictions that can be evaluated against existing datasets without waiting for the proposed experimental program. These are stated here as immediate falsifiability anchors.

First, in any published volatile reversal-learning or foraging dataset with simultaneous prefrontal population recordings, a low-dimensional population axis decoded without reference to behavioral outcomes should predict held-out switching cost and calibration error above difficulty, arousal, and attention controls. If no such axis is recoverable in datasets where the decoding pipeline has sufficient statistical power, the decodability claim fails in its current form.

Second, in any published multi-criterion fMRI or EEG dataset where participants switch between task objectives, criterion shifts should produce measurable RSA geometry change in prefrontal-insular regions above a pre-registered threshold. A result in which criterion shifts produce only global gain changes without geometric reorganization weakens H3 and H7 before any new data are collected.

Third, in any published reinforcement learning study comparing architectures with and without explicit recurrent bottleneck modules on volatile or reversal tasks, the bottleneck architecture should show superior switching cost and calibration performance relative to parameter-matched wider-attention baselines. Existing negative results in this comparison, if they survive strict

parameter matching, count against H6.

These predictions are offered not as retrospective fits but as prospective constraints. Any research group with access to the relevant datasets can evaluate them independently of the authors.

4.8 Criterion taxonomy

The criterion variable c is formally defined as a tuple

$$c = (L, \pi_{\text{ref}}, \tau), \tag{17}$$

where L is a criterion loss functional mapping a trajectory $\tau_{t:t+T}$ to a scalar, $L : \mathcal{T} \to \mathbb{R}$, π_{ref} is the reference policy against which control gain is measured, and τ is the horizon over which control gain is evaluated. This representation is general enough to subsume all criterion examples used in the manuscript. Reward prediction error corresponds to $L = -\mathbb{E}[\text{reward}]$, π_{ref} equal to the current reward expectation or baseline dopaminergic prediction, and τ near one step. Uncertainty minimization corresponds to L as posterior entropy, π_{ref} a uniform or exploration baseline, and τ variable. Anomaly detection corresponds to L as reconstruction or likelihood loss under a generative model. Long-horizon coherence corresponds to L as inconsistency penalty across extended action sequences and τ spanning multiple timesteps. The scientific value of this taxonomy is that it makes criterion manipulation concrete: two systems have the same criterion if and only if they share the same $(L, \pi_{\text{ref}}, \tau)$ tuple. Systems with identical architecture but different criterion tuples are therefore predicted to exhibit different relevance-manifold geometry even under identical sensory input, which is the falsifiable core of H3 and H7.

5 Methods

5.1 Model classes

The proposed artificial experiments compare at least three model classes.

Class M1 is a baseline neural model with attention but no explicit relevance state, such as a transformer policy network or recurrent controller with standard routing and hidden-state recurrence [16]. Class M2 augments the baseline with an explicit relevance module that computes r_t and reinjects it into latent state. Class M3 adds both relevance and metacognitive modules, allowing confidence and uncertainty readouts to be partially decoupled from relevance. Class M4 serves as a capacity-matched control in which attention width is increased without introducing an explicit relevance bottleneck.

Pre-registration commitment

The following must be locked before data collection begins in each experimental implementation: (1) functional forms of ϕ, A, Ψ, M, and G; (2) primary probe class - linear probes as primary mechanistic evidence, nonlinear probes as secondary sensitivity analysis; (3) intervention matching criterion - matched by norm, variance, and pilot-calibrated sensory-discrimination impairment; (4) primary outcome ratio - effect size for switching cost or calibration deficit under $\text{do}(r_t)$ must exceed effect size for basic sensory discrimination deficit by a pre-specified factor, default 1.5, to be confirmed by power analysis before unblinding; (5) null model specification - difficulty, arousal, attention, and reward regressors; (6) falsification thresholds - stated numerically before

data collection. Pre-registration is not a concession to open-science norms. It is what makes the falsification conditions binding.

5.2 Relevance-selection architecture

The relevance-selection pathway is defined in three stages:

1. candidate feature construction from external input, internal state, and context;
2. attention-based gating of candidate features;
3. criterion-conditioned estimation of expected control value and recurrent reinjection of the resulting relevance state.

In biological settings, the analogous mapping is not literal architectural identity but functional decomposition: sensory and contextual selection, relevance estimation under task and organismic constraints, and recurrent return into decision and control circuits.

5.3 Task design

Tasks are chosen to force dissociation among routing, relevance, confidence, and persistence.

The canonical task for both biological and artificial implementations is a volatile cue-conflict paradigm with the following structure: two competing cues presented simultaneously on each trial, one predicting criterion-relevant outcome under the current block, one salient but criterion-irrelevant; block length drawn from a geometric distribution with mean 20 trials, minimum 10; distractor stimuli present on 30% of trials; criterion switches unannounced; dependent variables are switching cost, trials to criterion after reversal, calibration error, gap between confidence proxy and accuracy, distractor susceptibility, performance decrement on distractor trials, and policy stability, action variance under constant latent contingencies. All other task variants are modifications of this canonical structure.

Selective attention task. Multiple cues compete, but only one predicts future control value.

Cue-shifting task. Previously relevant cues become irrelevant, testing switching thresholds and perseveration.

Uncertainty-report task. Agents must act and also report confidence or abstain.

Distraction-resistance task. High-salience distractors conflict with criterion-defined relevance.

Long-horizon consistency task. Short-term cues conflict with long-horizon objective, revealing whether relevance state stabilizes policy.

Cross-modal conflict task. Visual, auditory, or interoceptive streams carry inconsistent control information, testing criterion integration.

5.4 Observational measurements

Observational analysis proceeds in five layers.

First, decode candidate relevance variables from hidden states or population activity using linear probes, nonlinear probes, and state-space models. Second, compare decoded relevance with attention weights, confidence outputs, reward prediction error proxies, and prediction-error-like mismatch variables. Third, analyze temporal persistence of decoded relevance by autocorrelation, latent trajectory analysis, and recurrence statistics. Fourth, estimate latent manifold geometry under different criteria using representational similarity analysis and related geometric measures [27]. Fifth, evaluate whether decoded relevance predicts future switching, calibration, distractor resistance, and policy collapse beyond baseline hidden state and attention measures.

5.5 Causal interventions

Causal tests must distinguish relevance perturbation from generic damage.

In artificial systems, the primary manipulations are

$$\mathrm{do}(r_t \leftarrow 0), \qquad \mathrm{do}(r_t \leftarrow \mathrm{noise}), \qquad \mathrm{do}(r_t \leftarrow \hat{r}_t + \delta), \tag{18}$$

compared with matched perturbations on a_t, m_t, and broad latent subsets. Interventions are performed either at fixed timesteps, at threshold crossings of decoded relevance, or during specific task epochs such as post-cue/pre-response windows.

In biological systems, closed-loop decoding is used to estimate a relevance axis from neural population activity. Perturbation then targets candidate circuits at times when decoded relevance crosses predefined thresholds. The strongest implementation uses state-dependent stimulation or suppression rather than open-loop perturbation, because the latter is more easily confounded by global arousal, movement, or task epoch.

5.6 Control conditions

Control conditions include:

1. sham interventions;
2. nonspecific perturbation matched for injected variance or stimulation energy;
3. attention-gate perturbation without relevance perturbation;
4. metacognitive perturbation without relevance perturbation;
5. criterion-shuffled training or testing to dissociate criterion effects from architecture effects;
6. difficulty-matched conditions to ensure decoded relevance is not merely task load.

5.7 Outcome measures

Primary outcomes are accuracy, switching cost, policy stability, confidence calibration, abstention quality, distractor susceptibility, latency to criterion shift, long-horizon consistency, out-of-distribution robustness, and intervention sensitivity. Secondary outcomes are latent manifold separation, decoding accuracy for r_t, mutual information between r_t and future control, and dissociation metrics among a_t, r_t, and m_t.

5.8 Analysis strategy

The critical analysis principle is separation of observational and causal claims. Observational decoding establishes that relevance-like variables may exist. It does not establish necessity or mechanism. Causal claims require specific intervention effects and rival-model comparison. Statistical evaluation should therefore use hierarchical models or mixed-effects analyses across seeds, subjects, sessions, and training runs; robust interval estimates; and model comparison that penalizes added complexity. The theory is supported only if relevance perturbation explains behavior better than matched routing-only, confidence-only, reward-only, or predictive-error-only alternatives.

6 Experiments

6.1 Biological experiment

The primary biological implementation proposed here is a rodent volatile cue-conflict paradigm with large-scale neural recording and closed-loop perturbation. On each trial, the animal receives a sensory cue, a context cue, and a mild interoceptive or urgency manipulation. One cue predicts reward, another predicts threat or energetic cost, and contingencies shift unpredictably across blocks. The animal must choose, delay, or opt out.

At the level of a feasibility-validated pilot design, the species is adult male and female Long-Evans rats, or alternatively rhesus macaques if non-human primate recording is available to the research group. Recording is performed with chronic silicon probe arrays, for example Neuropixels 1.0 or equivalent [26], targeting prelimbic/infralimbic cortex, posterior parietal cortex, anterior insular cortex, and dorsomedial striatum simultaneously. The decoding pipeline uses linear dimensionality reduction, specifically principal component analysis followed by linear discriminant analysis or partial least squares, applied to multi-unit population vectors from held-out sessions with a minimum of 5-fold cross-validation. The candidate relevance axis is estimated on a discovery dataset using population variance structure alone - no behavioral outcome is used in axis estimation. Prediction of switching cost and calibration error is evaluated on a held-out confirmation dataset not used in axis definition. Discovery and confirmation datasets are separated by session, not trial, to prevent information leakage. Closed-loop stimulation is triggered when the decoded relevance-axis projection crosses 1.5 standard deviations above the within-session mean during a 200 ms post-cue window, prior to the response epoch. The stimulation modality is optogenetic suppression, using halorhodopsin or archaerhodopsin expressed via AAV, delivered as a 50 ms continuous pulse to the recording site showing the highest relevance-axis loading. The highest-loading site is treated as the primary perturbation target for the pilot study. In subsequent sessions, stimulation sites are varied systematically across candidate regions to establish causal mapping rather than assuming that loading magnitude identifies mechanism. Site is included as a fixed factor in the mixed-effects model. Statistical analysis uses a pre-registered mixed-effects logistic regression of switching probability on intervention condition, trial history, session, and animal, with random slopes for trial history and Bonferroni correction across the four intervention arms $\mathrm{do}(a_t)$, $\mathrm{do}(r_t)$, $\mathrm{do}(m_t)$, and $\mathrm{do}(h_t)$. If the relevance axis cannot be decoded above the difficulty-matched null distribution in at least 70% of sessions, the experiment is terminated and the negative result reported. Exact parameters such as pulse duration, expression system, and sample size remain subject to pre-registration before data collection begins.

Recordings include population activity from frontal, parietal, insular, and striatal circuits; trial-by-trial behavior; pupil diameter; respiratory state; and history variables. Candidate relevance axes are decoded using low-dimensional latent-variable models trained on held-out sessions. Closed-loop perturbation is triggered when the decoded relevance signal crosses a threshold during the post-cue/pre-choice epoch.

Predicted result: Why this task: volatile cue-conflict forces dissociation between attentional routing, which cue is currently salient, and relevance, which cue has control value under the current criterion. A low-dimensional population axis corresponding to relevance should forecast switching and calibration better than sensory evidence magnitude alone. State-dependent perturbation of that axis should selectively alter switching thresholds, context sensitivity, or distractor vulnerability. Failure modes include inability to separate relevance from arousal, weak decoding stability, or broad perturbation effects consistent with generic disruption rather than relevance-specific control.

6.1.1 Pilot stage

The biological study should begin as a pilot. The first goal is not to prove the full theory. It is to establish whether any candidate relevance axis can be decoded above null models that incorporate sensory salience, movement, vigilance, trial difficulty, and reward expectation.

Biological failure criterion: see falsification conditions in Sections 4.6 and 6.4.

6.1.2 Mechanistic rationale for area choice

Prefrontal cortex is justified because switching, rule maintenance, and control updating are expected there. Posterior parietal cortex is relevant for selective routing and evidence integration. Insular cortex is appropriate because bodily-state and salience-like variables are expected there. Dorsomedial striatum is relevant because criterion-linked action selection and policy transitions may be expressed there.

6.1.3 Controls

The biological study must include:

1. explicit movement regressors;
2. pupil- and respiration-based vigilance controls;
3. null decoders trained on shuffled criterion labels;
4. sensory-salience matched cue sets;
5. reward-matched threat and bodily-deviation conditions.

6.1.4 Trigger threshold

The closed-loop threshold should be justified from the pilot distribution of decoded relevance projections, not fixed dogmatically. A threshold around 1.5 standard deviations is a starting point, not a theoretical constant. Threshold choice should be cross-validated for stability and false-trigger rate.

6.1.5 Session count

The study should not imply that a single clean axis will simply emerge. A realistic claim is that dozens of sessions per animal, with repeated within-subject criterion shifts, may be needed before any stable relevance subspace can be estimated. Failure to recover such a subspace is informative and must be reported.

6.1.6 Minimal Protocol for Rodent Labs

To test for a candidate r_t-like subspace in your preparation:

Task: Volatile cue-conflict with interoceptive or urgency manipulation. Cue-reward contingencies shift unpredictably every 20–50 trials. Distractor stimuli appear on a subset of trials.

Recording: Chronic silicon probes, for example Neuropixels, targeting prelimbic/infralimbic cortex, posterior parietal cortex, anterior insular cortex, and dorsomedial striatum simultaneously.

Decoding: Linear dimensionality reduction, PCA plus LDA or PLS, applied to multi-unit population vectors. Cross-validate across held-out sessions. Define the candidate axis as the low-dimensional component that predicts held-out switching and calibration error, not used in axis definition, above a pre-registered threshold, for example likelihood-ratio test at $\alpha = 0.01$ with 5-fold cross-validation.

Perturbation: Closed-loop optogenetic suppression, for example halorhodopsin, of the recording site with highest axis loading. Trigger when decoded projection exceeds 1.5 standard deviations above within-session mean during a 200 ms post-cue window prior to response epoch. Pulse duration is 50 ms.

Failure criterion: see falsification conditions in Sections 4.6 and 6.4.

6.1.7 Human translation

Human methods can test whether an r_t-like subspace is decodable from prefrontal, insular, and parietal signals during criterion-manipulated tasks, whether that subspace predicts held-out switching, calibration, or distractor resistance above task difficulty, reaction time, confidence, arousal, stimulus energy, and generic activation controls, and whether criterion identity is decodable from the subspace in cross-validated holdout blocks. Human experiments therefore test convergence: whether a functional analogue of the relevance-feedback variable appears in the species of primary theoretical interest.

Human methods cannot directly implement $\mathrm{do}(r_t)$ on a decoded latent variable. fMRI, EEG, MEG, and TMS lack the spatial, temporal, and causal specificity required for direct intervention on a latent subspace. TMS disrupts a broad neural region, not a variable; lesion and pharmacological evidence is even coarser. The causal core of the framework is therefore tested in artificial systems and rodents, with human work serving as a convergent translation rather than a decisive test.

A positive human result means that the same operational structure is visible in humans: a low-dimensional, criterion-sensitive subspace predicts held-out control outcomes beyond difficulty, arousal, attention, and confidence controls, and coarse perturbation of its likely substrate produces a disproportionate control deficit relative to matched timing, site, and proxy controls. Such a result strengthens cross-species convergence but does not confirm the framework, because human methods cannot rule out task leakage, block structure, policy confounds, or regional

spillover. Sensory cortex involvement, if it survives regression of prefrontal activity, indicates broader broadcast but does not refute the framework; the framework only requires that the strongest decoding be in prefrontal-insular-striatal circuits, not that sensory decoding be absent.

A negative human result counts against the framework only when the same pipeline can decode other variables, including difficulty, confidence, or attention, from the same data; the study has sufficient statistical power; criterion manipulations change behavior in the predicted direction; and control conditions including sham, random timing, control site, and alternative decoded proxies are satisfied. Human null results that do not meet these conditions are uninformative. For the binding standards, see falsification conditions in Sections 4.6 and 6.4.

6.1.8 Neural signature predictions for candidate relevance subspaces

Status of neural signatures: The following are observational correlates of a candidate r_t-like subspace, not definitional criteria. They serve to distinguish candidate relevance from attention and metacognition when present, but failure of any specific signature, for example coherence or timescale difference, does not falsify the framework if the four primary criteria, decodability, persistence, criterion sensitivity, and causal efficacy, are satisfied. The central neural test remains the existence of a persistent, criterion-sensitive decoded subspace that predicts held-out control outcomes.

The framework does not claim that r_t corresponds to a single neuron, a narrow frequency band, or a stereotyped event-related potential. Instead, the candidate relevance subspace is a latent control variable distributed across populations and timescales. The predictions below concern neural correlates that should be observable if a candidate r_t-like subspace exists. They are not claims about mechanisms - they are observational signatures that distinguish candidate relevance from attention, metacognition, and generic latent state. The framework does not assume that any single signature is uniquely diagnostic; rather, the joint pattern across predictions constitutes evidence. These are observational signatures of candidate relevance subspaces and not direct evidence of mechanism; causal support for the framework comes only from the perturbation arms of the experimental program.

Firing rate dynamics: relatively slower than attention **Prediction:** In stable-criterion tasks, decoded relevance should decay more slowly than decoded attention on the same population, but this difference is expected to shrink when criteria shift rapidly. This is a relative prediction, not a fixed timescale law. The prediction is strongest in tasks with stable criterion blocks of at least 20–50 trials, and in rapidly shifting criterion tasks the timescale difference may be undetectable.

Rationale: Relevance integrates over recent evidence, criterion state, and task volatility. Attention can orient and reorient on each trial or even within a trial. Relevance should change more gradually unless the task demands rapid trial-by-trial criterion updates.

Measurable signature:

1. decoding a candidate r_t-like subspace, its autocorrelation should decay more slowly than that of decoded attention weights from the same neural population;

2. the difference in decay time constants, for example half-life in trials or seconds, should survive regressing out stimulus-locked responses, pupil-linked arousal, and movement.

Conditional qualification: In tasks with rapidly shifting criteria or no stable block structure,

even relevance may update quickly, and the decay difference from attention may become too small to detect.

Failure condition: If candidate relevance subspace dynamics are indistinguishable from attention dynamics in tasks where criteria are stable across blocks, the relative-slowness prediction fails.

Distinction from a_t: Attention is trial-local and access-oriented - it determines which features enter deeper processing on a short timescale. Relevance is criterion- and policy-oriented - it affects whether and when the system should switch strategies, persist, or recalibrate.

Timescales: multiple, with block-structured slow components when criteria are stable Prediction: The candidate relevance subspace should contain multiple timescales, including a slow component that aligns with criterion block boundaries or hazard-rate changes, but only when the task has such structure.

Rationale: Relevance estimation depends on the criterion c, which typically persists across blocks of trials. When the criterion changes, for example from reward maximization to uncertainty minimization, the geometry of the relevance subspace should reorganize over several trials. However, if criteria change on every trial, fast updates may dominate.

Measurable signature:

1. fitting a hierarchical or hidden Markov model to neural data, the inferred latent state that predicts switching should change more slowly than trial-by-trial stimulus fluctuations during stable criterion blocks;

2. a fast component, for example 1–3 trials, may still exist, tracking trial-specific control demands.

Failure condition: If all variance in the candidate relevance subspace operates on a single fast timescale indistinguishable from a_t or raw stimulus fluctuations, and the task includes stable criterion blocks of sufficient length, for example 30 or more trials, the multiple-timescale prediction fails. If the task lacks stable blocks, the prediction is not testable.

Distinction from confidence m_t: Confidence can update trial by trial based on outcome. Relevance should be more stable across consecutive trials unless the criterion or task structure changes.

Cross-area coherence: a candidate signature, not a general marker Prediction: Coherence is a secondary correlate that may covary with decoded relevance, but only after controlling for movement, arousal, and task engagement. If coherence fails to correlate with decoded relevance after these controls, the central decoding and perturbation claims remain the key tests of the framework.

Rationale: Recurrent relevance feedback may be supported by communication between regions involved in rule maintenance, bodily or criterion state, and policy selection. However, coherence can reflect many other processes.

Measurable signature:

1. coherence in the theta (4–8 Hz) or beta (13–30 Hz) band between prefrontal–insula, prefrontal–striatum, or insula–striatum may covary with periods when the decoded candidate relevance proxy is high, for example before a criterion-required switch, compared with periods when it is low;

2. this relationship must survive controls for movement, arousal, task engagement, and raw stimulus properties;

3. stronger control: the coherence–relevance correlation should be criterion-specific, so that the same physical cue produces different coherence patterns under different criteria even when movement, arousal, and engagement are matched.

Failure condition: If the coherence–relevance correlation disappears when movement, arousal, or engagement is regressed out, or if the same coherence pattern appears under all criteria for identical stimuli, the cross-area coherence prediction is not supported.

Important caveat: Cross-area coherence is nonspecific. The framework does not claim that theta or beta coherence is a "relevance rhythm." It claims only that if a candidate relevance subspace exists, it may be accompanied by coherence changes in these bands; if coherence fails to correlate with decoded relevance after movement, arousal, and task-engagement controls, the central decoding and perturbation claims remain the key tests of the framework.

If coherence does not correlate with decoded relevance after movement, arousal, and task-engagement controls, this does not weaken the framework's core claims, decodability and perturbation selectivity. It merely indicates that coherence is not a reliable signature in that preparation.

Region specificity: distributed, not exclusive to prefrontal–insular–striatal Prediction: The candidate relevance subspace should treat prefrontal–insular–striatal circuits as the primary candidate substrate, while allowing weaker or indirect relevance signals in sensory cortex via top-down feedback. A strong sensory decoding result that survives regression of prefrontal activity and stimulus-driven responses would weaken the anatomical localization claim but would not by itself refute the broader relevance-feedback architecture.

Rationale: Relevance is a control variable that operates after sensory selection and before action execution. However, feedback from prefrontal to sensory areas can modulate sensory representations according to relevance.

Measurable signature:

1. decoding should be stronger in prefrontal–insular–striatal circuits than in early sensory cortex after regression of stimulus-driven activity, and any sensory decoding that survives this regression and is not predictable from prefrontal activity alone would require explanation beyond the current architecture;

2. what remains in sensory areas should usually be indirect, for example predictable from prefrontal activity, and should weaken when prefrontal activity is statistically controlled.

Failure condition: If candidate r_t is equally decodable from early sensory cortex as from prefrontal cortex, and this sensory decoding survives controls and is not predictable from prefrontal activity, the region-specificity prediction weakens, but the broader relevance-feedback architecture is not refuted, because relevance could still be computed elsewhere and broadcast.

Distinction from a_t: Attention is decodable from parietal and early sensory regions, where top-down modulation is well established. Candidate relevance is predicted to be most robust in anterior circuits, with sensory representations being weaker and indirect.

Sensory cortex involvement, if it survives regression of prefrontal activity, suggests broader broadcast but does not refute the framework. The framework only requires that the strongest decoding be in prefrontal-insular-striatal circuits, not that sensory decoding be absent.

Criterion shift reorganization: geometric change detectable with RSA Prediction: When criterion c changes, for example from reward to uncertainty minimization, criterion shifts should produce measurable changes in representational geometry, as operationalized by RSA dissimilarity matrices or a pre-registered mixed-effects model of the form Dissimilarity $\sim$ CriterionPair + StimulusIdentity + (1|Subject/Session), with the specific threshold for what constitutes meaningful reorganization stated before data collection. This is a specific claim about representational similarity structure, not just a change in mean firing rates.

Rationale: Different criteria define different mappings from signals to control gain. The same physical cue may be highly relevant under one criterion and irrelevant under another. This should be reflected in a changed representational geometry.

Measurable signature:

1. using representational similarity analysis (RSA), compute the dissimilarity matrix, for example 1 - correlation or Euclidean distance, of neural population states for trials under different criteria [27];

2. within-criterion dissimilarity should be lower than between-criterion dissimilarity for the same stimuli, after controlling for overall firing-rate differences, stimulus identity, and block-order effects;

3. what counts as genuine reorganization is a significant interaction in the pre-registered model that survives cross-validation and exceeds the pre-specified threshold.

Failure condition: If criterion shifts produce only a global scaling of neural activity, without geometric reorganization detectable by RSA after the controls above, the criterion-specificity prediction fails.

Distinction from a_t: Attention can change gain without changing geometry. Candidate relevance is predicted to require geometric reorganization because it changes the meaning of control relevance, not just the weight on features.

Perturbation effects: selective relative deficit, not absolute sensory preservation Prediction: Optogenetic suppression or activation of the candidate r_t substrate should produce a selectively disproportionate impairment in switching cost, calibration error, or distractor resistance relative to matched do(a_t), do(m_t), and do(h_t) interventions, where disproportionate means the ratio of switching or calibration deficit to basic sensory discrimination deficit is larger for do(r_t) than for the other arms. Basic sensory discrimination should be relatively more preserved than switching and calibration outcomes, though some spillover is expected because the targeted circuits are embedded in broader networks; a result in which sensory discrimination degrades proportionally to switching cost would count against the framework, while a result in which sensory discrimination degrades less than switching cost would be consistent with it even if not fully preserved. A reasonable pre-registration threshold for acceptable spillover would be that the effect size for switching cost or calibration deficit under do(r_t) is at least 1.5 times larger than the effect size for basic sensory discrimination deficit under the same intervention, with this ratio pre-specified before unblinding.

Rationale: r_t operates downstream of sensory selection and upstream of policy. Disrupting its neural substrate should impair when to switch and how calibrated the system is more than what the stimulus is. However, spillover effects are expected because neural circuits are interconnected.

Measurable signature:

1. following do(r_t), for example optogenetic suppression of a region identified as part of the candidate relevance subspace, there should be a large effect on switching cost, measured as more trials to adapt after rule reversal;
2. there should be a large effect on calibration, measured as mismatch between confidence proxies and accuracy;
3. there should be a large effect on distractor susceptibility;
4. there may be a smaller effect on simple discrimination accuracy for highly salient, non-switching trials;
5. reaction-time increases should not fully explain the disproportionate deficits;
6. because optogenetic suppression affects all processes in the targeted region and its efferents, some sensory or motor impairment is expected. The claim is relative: the ratio of switching or calibration deficit to sensory deficit should be larger for relevance-targeted perturbation than for attention-targeted perturbation.

Failure condition: If do(r_t) produces a globally proportional performance collapse, so that all metrics degrade equally including simple sensory discrimination, with no metric showing disproportionate impairment, or if the effect is zero, the selective-deficit prediction fails.

Distinction from do(a_t)**:** Attention perturbation should impair early feature selection disproportionately, for example producing a larger effect on distractor filtering than on switching. r_t perturbation should impair later control decisions disproportionately, for example producing a larger effect on switching cost than on distractor filtering.

Summary table: Neural signatures of candidate r_t versus a_t versus m_t *H9 note:* The criterion-invariant subspace z^* is predicted to be decodable from the same prefrontal-insular-striatal circuits as r_t, with the distinguishing feature that its geometry does not reorganize under criterion change.

Note on the table: Entries describe candidate signatures of the respective latent variables. The table does not imply that any variable has been isolated in real neural data - only what one should look for.

What would falsify neural predictions - and which failure is most damaging The neural component of the framework would be substantially weakened if any of the following occur. They are not equally informative.

Most damaging, central to the framework: If no candidate relevance subspace predicts held-out control outcomes above difficulty, arousal, value, and attention controls, and this failure persists across at least two independent task paradigms with sufficient statistical power, the strong neural claim is falsified. To prevent a single underpowered null result from appearing to falsify the framework, and to prevent a single positive result from appearing to confirm it, the minimum evidentiary standard for both confirmation and falsification of the neural claim is replication across at least two independent task paradigms, two species or system types, or one biological and one artificial implementation.

Moderately damaging, requiring revision but not abandonment:

1. criterion shifts produce only global gain changes, not geometric reorganization detectable by RSA, and the task includes stable criterion blocks;

2. perturbation produces no selective deficit pattern, so that all metrics degrade proportionally or not at all;

3. decoded candidate relevance has a timescale indistinguishable from attention in tasks with stable criterion blocks.

Least damaging, possibly reflecting implementation details rather than the core theory:

1. cross-area coherence does not correlate with the decoded proxy, since coherence is only a secondary signature;

2. region specificity differs from predictions, for example strong sensory encoding survives controls, because the framework can accommodate this by allowing top-down modulation to embed relevance widely;

3. multiple timescales are not observed in tasks without stable block structure, because the prediction is conditional.

Summary: The central neural failure condition is failure to decode any candidate relevance subspace that predicts held-out control outcomes above difficulty, arousal, value, and attention controls across the minimum replication standard; see falsification conditions in Sections 4.6 and 6.4. Secondary predictions, including timescales, coherence, geometry, and perturbation selectivity, serve to distinguish candidate relevance from confounds but are not individually fatal if they fail, provided decoding succeeds.

Neural signatures versus neural mechanisms **Important caveat:** The predictions above concern observational signatures of candidate relevance subspaces. They do not claim that the recorded activity is the mechanism of relevance estimation. A region could be necessary for relevance-dependent behavior without encoding r_t in a decodable form, for example because it acts as a relay or a necessary modulator. Conversely, decodability does not imply causality. The framework's causal claims rest on the perturbation experiments in this section, not on neural signatures alone. The signatures described here are tools for identifying candidate substrates to target for perturbation, not final evidence for the theory.

6.2 AI experiment

The primary artificial implementation uses a parameter-matched family of sequential agents trained on partially observable control tasks with volatile contingencies, distractors, and long-horizon objectives. M1–M4 from Table 2 are trained under either reward-only criteria or hybrid criteria combining reward, uncertainty, anomaly, and coherence.

For reproducibility, the proof-of-concept environment is a 4×4 partially observable gridworld with block-structured reward switching and salient distractor states. The reward function is $R(s, a, c) = +1$ for reaching the criterion-defined goal state, -0.1 per step, and -1 for entering a distractor state. Criterion switches follow a geometric holding-time distribution with mean block length of 20 trials and a minimum of 10 trials, ensuring sufficient within-block data for relevance decoding while requiring genuine switching behavior. The bottleneck sweep runs $\dim(r_t) \in \{1, 2, 4, 8, 16, 32\}$ under matched random seeds and identical reinforcement-learning objectives across M1–M4.

To extend beyond proof-of-concept, a second environment is required with greater state complexity and longer horizon structure. The recommended candidate is a partially observable variant of MiniGrid or BabyAI with at least 64 states, multi-step criterion satisfaction, and embedded distractor objects. Results from the 4×4 gridworld establish the intervention logic; results from the extended environment establish generalization. Claims about awareness-like control in AI systems require both.

For a concrete proof-of-concept implementation, the baseline agent can be instantiated as a 6-layer causal transformer controller with recurrent memory, matched hidden size across conditions, and an explicit relevance bottleneck in M2 and M3. The bottleneck dimension k is swept over a small range, for example $k \in \{1, 2, 4, 8, 16, 32, \dim(h_t)/2\}$, to test H8 directly under matched optimization settings. Training uses identical task curricula, matched random seeds, and the same reinforcement-learning objective across classes, with the only architectural differences being the relevance bottleneck and metacognitive head. Probes are fit on held-out runs. Targeted interventions clamp, scramble, or adversarially perturb r_t during the same post-cue/pre-response window used in the biological design so that time-resolved phenotypes can be compared across domains.

The artificial perturbation arms, $\mathrm{do}(a_t)$, $\mathrm{do}(r_t)$, $\mathrm{do}(m_t)$, and $\mathrm{do}(h_t)$, follow the same intervention logic and timeline in Figure 2 as the biological study, enabling cross-domain comparison.

Measurements include task performance, calibration, hidden-state probes, counterfactual clamping of r_t, representational geometry, robustness under distribution shift, and criterion-transfer behavior. The main analysis asks whether an explicit recurrent relevance module yields a decodable, low-dimensional state that predicts control outcomes beyond attention and whether direct intervention on that state yields specific failure signatures.

Predicted result: relevance-augmented models should improve calibration, strategic switching, and robustness while producing latent relevance states that are not reducible to attention maps. Failure modes include disappearance of gains after strict scale matching, collapse of relevance probes onto difficulty measures, or intervention effects no more specific than general hidden-state corruption.

Gridworld is a proof-of-concept, not a proof-of-generalization. Scaling to large language models or real-world robotics is a separate engineering project. The framework provides the causal logic and falsification criteria; scaling is implementation work for subsequent papers. A framework that only worked on large-scale systems could not be tested efficiently.

6.2.1 Minimal Implementation for AI Researchers

A minimal PyTorch-style implementation of M2, the relevance-augmented model, adds three conceptual lines to a standard transformer policy network:

```
# Assume h_t is the latent state from the transformer (batch, dim_h)
relevance_logits = W_r @ h_t           # project to candidate relevance space
r_t = compress(relevance_logits, dim=r_dim)  # bottleneck, e.g., linear or softmax
h_t_updated = h_t + W_return @ r_t     # reinject into latent dynamics before policy
```

A reproducible implementation will be released concurrent with the first empirical results from this framework. The core prediction is that M2 and M3 will outperform M1 and M4 on switching cost and calibration error by a margin exceeding the 95% confidence interval of the baseline, for example $\geq 10\%$ relative improvement, and that $\mathrm{do}(r_t)$ clamping will produce a

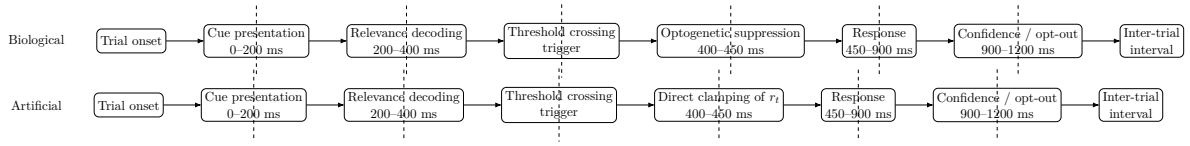

Figure 2: Intervention timeline for the causal perturbation study. Top row: biological implementation using closed-loop optogenetic suppression triggered at relevance-axis threshold crossing. Bottom row: artificial implementation using direct clamping of the r_t state vector. Both rows align on the same task epoch structure to enable cross-domain comparison. The 200 ms window is an implementation starting point to be calibrated from pilot data, not a theoretically fixed parameter.

selectively disproportionate impairment in switching or calibration relative to $\mathrm{do}(a_t)$, $\mathrm{do}(m_t)$, and $\mathrm{do}(h_t)$.

6.3 Cross-domain comparison

A cross-domain experiment uses homologous evidence-accumulation and cue-reversal tasks in animals and AI agents. The goal is not to claim phenomenological equivalence. It is to ask whether criterion-conditioned relevance produces parallel signatures: decodable low-dimensional state, persistence across timesteps, dissociation from attention, and selective perturbation sensitivity.

Predicted result: both domains should exhibit criterion-dependent control manifolds, but biological systems should weigh threat, bodily deviation, and temporal urgency more strongly than reward-matched artificial systems unless the latter are explicitly trained to do so. Failure modes include no common dynamical signature or only trivial similarities caused by surface task design rather than relevance architecture.

6.4 Causal perturbation study

The causal perturbation study is a causal perturbation program with four arms:

$$\mathrm{do}(a_t), \qquad \mathrm{do}(r_t), \qquad \mathrm{do}(m_t), \qquad \mathrm{do}(h_t). \tag{19}$$

Predicted result: $\mathrm{do}(r_t)$ should produce a selectively disproportionate impairment in switching cost, calibration error, or distractor resistance relative to matched $\mathrm{do}(a_t)$, $\mathrm{do}(m_t)$, and $\mathrm{do}(h_t)$ interventions, where disproportionate means the ratio of switching or calibration deficit to basic sensory discrimination deficit is larger for $\mathrm{do}(r_t)$ than for the other arms. Basic sensory discrimination should be relatively more preserved than switching and calibration outcomes, though some spillover is expected because the targeted circuits are embedded in broader networks; a result in which sensory discrimination degrades proportionally to switching cost would count against the framework, while a result in which sensory discrimination degrades less than switching cost would be consistent with it even if not fully preserved. $\mathrm{do}(a_t)$ should primarily affect access and feature selection. $\mathrm{do}(m_t)$ should affect report or explicit confidence more than core control. $\mathrm{do}(h_t)$ should produce diffuse impairment. Distinguishing these proportional phenotypes is the falsifiability center of the theory.

7 Critical Rival Theories

7.1 Attention as pure routing

Pure routing theories identify attention with selective weighting: information becomes behaviorally influential because it receives computational priority through gain, routing, or access. This view explains how signals enter processing, but it does not explain why a selected signal becomes a persistent recurrent control variable that changes switching, calibration, or distractor resistance after selection has already occurred. In the four-arm study, pure routing predicts that $\mathrm{do}(r_t)$ approximates $\mathrm{do}(a_t)$, because relevance is not a separate recurrent state. The present framework predicts selectively disproportionate impairment in switching, calibration, or distractor resistance after $\mathrm{do}(r_t)$ relative to matched $\mathrm{do}(a_t)$, $\mathrm{do}(m_t)$, and $\mathrm{do}(h_t)$. The discriminating result is a larger control-deficit-to-sensory-deficit ratio under $\mathrm{do}(r_t)$ than under matched attention-gate perturbation. Work showing that attention weights are not reliable explanations strengthens this separation in artificial systems [17, 18].

7.2 Awareness as global workspace access

Global workspace and global neuronal workspace theories explain conscious access as large-scale broadcast or ignition, especially under report conditions [4, 5]. Their strongest version holds that contents become conscious when they are globally available to multiple systems, including report, memory, and flexible control. The present framework positions r_t as a pre-broadcast relevance selector - the variable that determines what enters global ignition, not the ignition event itself. This is a specific architectural claim: r_t should be decodable and perturbable before broadcast ignition occurs. A result in which $\mathrm{do}(r_t)$ impairs control only after full ignition is established weakens this positioning. In the four-arm study, global workspace theory predicts that relevance effects depend on broadcast ignition; the present framework predicts selectively disproportionate impairment from $\mathrm{do}(r_t)$ even when full global ignition is not established. The discriminating result is an early decoded relevance subspace whose perturbation changes switching or calibration before global broadcast signatures are present.

7.3 Awareness as metacognition

Higher-order and metacognitive theories argue that awareness depends on representations of one's own mental states or self-evaluative monitoring [8, 6, 7]. Their strongest version predicts that awareness tracks confidence, uncertainty, or higher-order access to first-order states. The present framework treats metacognition as one readout family within a broader relevance-feedback architecture rather than as the whole mechanism. In the four-arm study, metacognitive theories predict that $\mathrm{do}(m_t)$ approximates $\mathrm{do}(r_t)$, because awareness-relevant disruption should follow self-evaluative monitoring. The present framework predicts dissociation: $\mathrm{do}(m_t)$ affects confidence and report more strongly, whereas $\mathrm{do}(r_t)$ produces disproportionate impairment in switching, calibration, or distractor resistance. The discriminating result is preserved or partially preserved confidence readout alongside a larger control deficit after $\mathrm{do}(r_t)$, or confidence disruption without the same control phenotype after $\mathrm{do}(m_t)$.

7.4 Relevance as reward prediction

Reward prediction error theory explains adaptive learning through discrepancies between expected and obtained reward [13, 14, 12, 15]. Its strongest version reduces relevance-driven control to

value update or reward-prediction residuals. The present framework treats reward prediction error as one special case of criterion-bound relevance rather than a universal account, because threat, anomaly, homeostatic deviation, urgency, and coherence failure can dominate control before reward is resolved. In the four-arm study, reward-prediction accounts predict that $\mathrm{do}(r_t)$ effects are explained by reward-prediction residuals under reward-matched controls. The present framework predicts disproportionate relevance-perturbation effects even when expected reward is held constant and hazard, interoceptive load, anomaly, or coherence is varied. The discriminating result is a decoded relevance state that tracks those manipulations after residualizing out estimated reward prediction error.

7.5 Relevance as predictive processing

Predictive coding explains hierarchical error propagation and feedback [9, 10]. Its strongest version treats perceptual and control updating as the management of prediction error across hierarchical levels. Active inference specifies expected free energy and precision-weighting as computable quantities. The present framework does not compete with these as general brain theories. It makes a narrower, differently testable claim: the relevance state must be decodable as a separable low-dimensional variable, geometrically reorganized by criterion change, and selectively interruptible with a disproportionate behavioral phenotype. Precision-weighted free-energy minimization does not require any of these properties, and active inference in its current form does not commit to a selectively perturbable r_t subspace. These are additive requirements, not contradictions. In the four-arm study, predictive processing predicts that $\mathrm{do}(r_t)$ effects track prediction-error magnitude; the present framework predicts that some low-error states remain highly relevant because of strategic, bodily, or temporal consequence. The discriminating result is disproportionate $\mathrm{do}(r_t)$ impairment in conditions where prediction-error magnitude is matched or low but criterion-defined control value is high.

7.6 Relevance as free-energy minimization

The free-energy principle offers an unusually broad account of action, perception, and learning [11]. Its strongest version can redescribe adaptive behavior as variational updating that minimizes free energy. Active inference specifies expected free energy and precision-weighting as computable quantities. The present framework does not compete with these as general brain theories. It makes a narrower, differently testable claim: the relevance state must be decodable as a separable low-dimensional variable, geometrically reorganized by criterion change, and selectively interruptible with a disproportionate behavioral phenotype. Precision-weighted free-energy minimization does not require any of these properties, and active inference in its current form does not commit to a selectively perturbable r_t subspace. These are additive requirements, not contradictions. Active inference, as a specific implementation of the free-energy principle, already distinguishes epistemic value from pragmatic value and uses precision-weighting to implement criterion-sensitive relevance [28]. The four-arm perturbation study provides a test that active inference in its current form does not specify, because it does not commit to a decodable and selectively interruptible relevance subspace. In the four-arm study, free-energy accounts predict broad variational-update disruption rather than a selectively perturbable subspace distinct from global updating; the present framework predicts a low-dimensional r_t variable with criterion-sensitive geometry and disproportionate perturbation effects. The discriminating result is selective interruption of the decoded relevance variable without reducing the finding to broad free-energy or precision-weighting language.

7.7 Awareness as integrated information

Integrated information theory makes strong claims about intrinsic causal structure and the conditions under which experience exists [19, 20]. Its strongest version predicts that conscious status depends on system-intrinsic integrated causal organization rather than on task-specific control performance alone. The present framework is less metaphysically ambitious and more directly experimental: it predicts how different criteria reshape internal state geometry and control under intervention. In the four-arm study, integrated information theory predicts perturbation effects that scale with integrated information rather than criterion geometry; the present framework predicts criterion-sensitive reorganization of r_t and selectively disproportionate control impairment under $\mathrm{do}(r_t)$. The discriminating result is criterion-dependent relevance geometry and perturbation specificity that changes with c even when broad system integration is held approximately constant.

8 Discussion

The deepest claim of the paper is not that consciousness is solved, nor that a single latent variable captures all of conscious life. The claim is narrower and stronger: awareness-like control can be studied as the causal role of criterion-sensitive relevance feedback in ongoing computation. This shifts the discourse from verbal analogy to measurable architecture.

Several hidden assumptions must be stated openly. First, the framework assumes that latent control variables can be usefully decomposed into routing, relevance, and metacognitive components. This may fail in highly entangled systems. Second, it assumes that interventions can target relevance with enough specificity to produce discriminable phenotypes. This is experimentally demanding. Third, it assumes that criterion c can be stated explicitly enough in artificial systems and approximated well enough in biological tasks to support comparison. Fourth, it assumes that reportability is neither necessary nor sufficient for awareness-like relevance, which is defensible but contestable.

The strongest novelty opportunity lies in treating *criterion* as a first-class scientific object. Most theories ask whether attention, broadcast, higher-order representation, predictive error, or integrated causal structure is sufficient. This framework asks a different question: what class of criterion determines which internal states become recurrently control-dominant? Once stated this way, biological and artificial systems become directly comparable without pretending that they are psychologically identical. Biological systems are expected to exhibit priors for threat, homeostasis, and social urgency. Artificial systems can be engineered to prioritize other variables. That means awareness-like control is not a binary metaphysical property in this framework. It is a family of criterion-dependent recurrent control regimes.

This perspective has immediate implications for AI. A model may have attention without relevance feedback, confidence without robust control, or policy without explicit internal relevance state. Conversely, a model with a clearly decodable, perturbable relevance variable could exhibit awareness-like control signatures without implying full human-like consciousness. This distinction is scientifically and ethically useful because it prevents both over-attribution and under-theorization. More precisely: the framework provides a principled lower bound for when awareness-like control attribution is scientifically warranted, and a principled upper bound for when it is not — and both bounds matter for how artificial systems should be evaluated, governed, and assigned moral consideration in contexts where such questions are no longer merely theoretical.

H9 raises a question the present framework can gesture toward but cannot resolve empirically. If criterion-invariant representations exist because the conditions that shaped biological systems, physical, energetic, temporal, imposed common structure across criteria, then the deepest invariants in z^* would reflect not the contingent history of one evolutionary lineage but structural features of any causal information-processing system operating within this universe's physical and mathematical constraints. Representations of temporal order, causal structure, uncertainty, and information content may be criterion-invariant not merely because evolution selected for them but because any system that accurately tracks a world with these properties must represent them. This is a speculative horizon, not a finding. The experimental program proposed here tests the biological and artificial z^* directly. Whether the deepest layers of z^* reflect mathematical necessity rather than evolutionary contingency is a question for subsequent theoretical and formal work, not for the perturbation studies proposed here.

9 Limitations

The following limitations are stated here in full. Key limitations, including underspecified functions, non-computable counterfactual relevance, and non-uniqueness, are also noted in Sections 3.4, 4.1, and 4.4.

This paper is a theory-and-experiment framework rather than a completed empirical report. Its central claims require the proposed perturbation studies.

The formal architecture specifies the causal structure and intervention logic of the framework but does not constrain the functional form of its component functions, ϕ, A, Ψ, M, and G, which must be instantiated and pre-registered within each experimental implementation. This underspecification means that the framework can be made consistent with a wide range of results through post-hoc choice of components, and the primary safeguard against this is pre-registration of model classes and analysis pipelines before data collection.

The most direct form of this objection is: if the functional forms of ϕ, A, Ψ, M, and G are left open, any result can be made consistent with the framework post-hoc. This objection is incorrect, and the reason it is incorrect is the intervention ratio. No choice of functional forms can produce the predicted do(r_t) phenotype — selectively disproportionate impairment in switching and calibration relative to matched do(a_t), do(m_t), and do(h_t) — if relevance feedback is not a real separable control variable in the system being tested. The four-arm ratio is architecture-agnostic precisely because it tests causal separability, not functional form. A system in which a_t, r_t, and m_t are fully entangled will produce proportionally uniform impairment across all four arms regardless of the functional forms chosen, and that result falsifies the framework regardless of what post-hoc fitting might suggest. The falsifiability of the framework therefore rests on the intervention logic, not on the specification of component functions. Pre-registration of functional forms before data collection is the procedural safeguard; the intervention ratio is the logical safeguard. Both are required. Neither alone is sufficient.

The counterfactual control-gain definition of relevance in Equation 5 is not directly computable in most real systems. The approximation cascade in Section 3.4 addresses this partially, but the gap between the theoretical definition and any tractable estimator remains a limitation that future work must close more completely.

The four identification criteria in Section 4.4 do not guarantee uniqueness. Multiple latent variables could in principle satisfy decodability, temporal persistence, criterion sensitivity, and causal efficacy simultaneously. The framework identifies candidate relevance dimensions subject

to further causal discrimination, not a provably unique decomposition.

The cross-area coherence prediction is a secondary correlate and not a central test. If coherence fails to survive controls for movement, arousal, and task engagement, the framework's core claims about decodability and causal perturbation specificity are unaffected.

The gridworld proof-of-concept environment for the AI experiment is deliberately simple. Generalization of relevance-attention dissociations to large-scale architectures such as large language models or real-world agents is not guaranteed and requires separate investigation.

The framework does not resolve phenomenology. It addresses awareness-like control in a functional and causal sense.

The decomposition into a_t, r_t, and m_t may be only approximate in real brains and in many network classes. In some systems, these quantities may exist only as partially separable directions in a shared latent manifold. The four identification criteria in Section 4.4, decodability, temporal persistence, criterion sensitivity, and causal efficacy, are designed to provide experimentally useful dissociation even in systems where a_t, r_t, and m_t are not architecturally isolated, and they constitute the operative test of separability rather than an assumption of it.

Biological relevance criteria are only partially observable and may shift across development, training, bodily state, and context. Artificial relevance can be engineered, but poorly chosen criteria may generate pathological salience landscapes, reward hacking, brittle internal control, or misleading confidence signals.

The framework lacks a complete substrate-independent scaling law that would allow direct quantitative comparison of awareness-like control across radically different architectures. At present, comparison depends on task structure, latent geometry, and intervention phenotype rather than on one universal scalar.

H9, criterion-invariant structure, raises the possibility that the deepest invariants in z^* reflect mathematical or physical necessity rather than evolutionary contingency. This is a speculative horizon that the proposed experiments cannot resolve and that is left for subsequent theoretical work.

Phenomenology - subjective experience, qualia, what it is like to be a system - is outside the scope of this framework. The framework addresses functional control signatures. This is not evasion: it is scope discipline. A framework that claimed to resolve phenomenology while proposing causal perturbation studies would be making promises its methods cannot keep. The present framework makes only the promises its methods can test. The framework makes no claims about inner experience. It claims that functional control properties (switching, calibration, distractor resistance, policy stability) are sufficient to define and test awareness-like control. If a system passes the four-arm perturbation study, it behaves as if it has awareness-like control. Whether it "actually" experiences anything is a separate question that this framework does not answer - and does not need to answer to be useful.

Finally, some rival theories may be empirically close enough that only very carefully designed interventions will separate them. If those interventions cannot be performed with sufficient specificity, the strongest version of the theory remains unverified.

10 Conclusion

Biological intelligence does not process all available data equally. It selects what matters under pressures tied to control, uncertainty, and survival. This paper has argued that awareness can be studied as a causal regime of relevance selection rather than as a synonym for attention

or a purely philosophical label. In the proposed framework, attention gates candidate signals, relevance estimates their expected control value under a criterion, internal feedback returns a compact relevance state into the system, and relevance-feedback state transitions occur when that recurrent state alters subsequent processing in a specific and testable way.

This account distinguishes attention from awareness, confidence from relevance, reward prediction from broader control salience, predictive error from criterion-bound importance, and broad formal frameworks from experimentally localizable control variables. Its scientific value lies in what it forbids as much as in what it allows: it forbids calling a system awareness-like on the basis of routing alone, report alone, or correlation alone. It requires decodability, criterion dependence, and causal intervention.

If that program succeeds, it offers a comparative science of awareness-like control across brains and machines. If it fails, the failure will still be informative, because it will show that attention, metacognition, predictive processing, reward prediction, free-energy minimization, or global access can explain more than this framework grants them. Either outcome is empirically valuable.

References

[1] Posner, M. I., and Petersen, S. E. The attention system of the human brain. *Annual Review of Neuroscience* **13**, 25–42 (1990). doi:10.1146/annurev.ne.13.030190.000325.

[2] Petersen, S. E., and Posner, M. I. The attention system of the human brain: 20 years after. *Annual Review of Neuroscience* **35**, 73–89 (2012). doi:10.1146/annurev-neuro-062111-150525.

[3] Knudsen, E. I. Fundamental components of attention. *Annual Review of Neuroscience* **30**, 57–78 (2007). doi:10.1146/annurev.neuro.30.051606.094256.

[4] Dehaene, S., and Changeux, J.-P. Experimental and theoretical approaches to conscious processing. *Neuron* **70**(2), 200–227 (2011). doi:10.1016/j.neuron.2011.03.018.

[5] Mashour, G. A., Roelfsema, P., Changeux, J.-P., and Dehaene, S. Conscious processing and the global neuronal workspace hypothesis. *Neuron* **105**(5), 776–798 (2020). doi:10.1016/j.neuron.2020.01.026.

[6] Fleming, S. M., and Dolan, R. J. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B* **367**(1594), 1338–1349 (2012). doi:10.1098/rstb.2011.0417.

[7] Yeung, N., and Summerfield, C. Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B* **367**(1594), 1310–1321 (2012). doi:10.1098/rstb.2011.0416.

[8] Lau, H., and Rosenthal, D. Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences* **15**(8), 365–373 (2011). doi:10.1016/j.tics.2011.05.009.

[9] Rao, R. P. N., and Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* **2**(1), 79–87 (1999). doi:10.1038/4580.

[10] Friston, K. A theory of cortical responses. *Philosophical Transactions of the Royal Society B* **360**(1456), 815–836 (2005). doi:10.1098/rstb.2005.1622.

[11] Friston, K. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* **11**, 127–138 (2010). doi:10.1038/nrn2787.

[12] Schultz, W. Dopamine reward prediction-error signalling: a two-component response. *Nature Reviews Neuroscience* **17**, 183–195 (2016). doi:10.1038/nrn.2015.26.

[13] Waelti, P., Dickinson, A., and Schultz, W. Dopamine responses comply with basic assumptions of formal learning theory. *Nature* **412**, 43–48 (2001). doi:10.1038/35083500.

[14] Eshel, N., Tian, J., Bukwich, M., and Uchida, N. Dopamine neurons share common response function for reward prediction error. *Nature Neuroscience* **19**, 479–486 (2016). doi:10.1038/nn.4239.

[15] Starkweather, C. K., Babayan, B. M., Uchida, N., and Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nature Neuroscience* **20**(4), 581–589 (2017). doi:10.1038/nn.4520.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, 5998–6008 (2017).

[17] Jain, S., and Wallace, B. C. Attention is not explanation. In *Proceedings of NAACL-HLT 2019*, 3543–3556 (2019). doi:10.18653/v1/N19-1357.

[18] Wiegreffe, S., and Pinter, Y. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP 2019*, 11–20 (2019). doi:10.18653/v1/D19-1002.

[19] Tononi, G. An information integration theory of consciousness. *BMC Neuroscience* **5**, 42 (2004). doi:10.1186/1471-2202-5-42.

[20] Oizumi, M., Albantakis, L., and Tononi, G. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology* **10**(5), e1003588 (2014). doi:10.1371/journal.pcbi.1003588.

[21] Lau, H. C., and Passingham, R. E. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences of the USA* **103**(49), 18763–18768 (2006). doi:10.1073/pnas.0607716103.

[22] Vorberg, D., Mattler, U., Heinecke, A., Schmidt, T., and Schwarzbach, J. Different time courses for visual perception and action priming. *Proceedings of the National Academy of Sciences of the USA* **100**(10), 6275–6280 (2003). doi:10.1073/pnas.0931489100.

[23] Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, 368–377 (1999).

[24] Tishby, N., and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop*, 1–5 (2015). doi:10.1109/ITW.2015.7133169.

[25] Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. In *Proceedings of the International Conference on Learning Representations* (2017). arXiv:1612.00410.

[26] Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., Lee, A. K., Anastassiou, C. A., Andrei, A., Aydın, Ç., Barbic, M., Blanche, T. J., Bonin, V., Couto, J., Dutta, B., Gratiy, S. L., Gutnisky, D. A., Häusser, M., Karsh, B., Ledochowitsch, P., Lopez, C. M., Mitelut, C., Musa, S., Okun, M., Pachitariu, M., Putzeys, J., Rich, P. D., Rossant, C., Sun, W.-L., Svoboda, K., Carandini, M., Harris, K. D., Koch, C., O'Keefe, J., and Harris, T. D. Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**(7679), 232–236 (2017). doi:10.1038/nature24636.

[27] Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* **2**, 4 (2008). doi:10.3389/neuro.06.004.2008.

[28] Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. Active inference: a process theory. *Neural Computation* **29**(1), 1–49 (2017). doi:10.1162/NECO_a_00912.

[29] Schmidhuber, J. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development* **2**(3), 230–247 (2010). doi:10.1109/TAMD.2010.2056368.

[30] Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, 2778–2787 (2017).

[31] Klyubin, A. S., Polani, D., and Nehaniv, C. L. Empowerment: a universal agent-centric measure of control. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, 128–135 (2005). doi:10.1109/CEC.2005.1554676.

[32] Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., and Abbeel, P. VIME: Variational Information Maximizing Exploration. In *Advances in Neural Information Processing Systems 29*, 1109–1117 (2016).

Table 1: Operational separation of the core constructs used in the framework. These definitions are operative throughout the manuscript. Where the paper uses any of these terms, the meaning is the one given here, not the colloquial or field-specific usage.

Construct	**Operational definition**	**Primary observable signature**
Attention	A routing or gating process that changes which candidate features receive computational priority.	Selection weights, routing masks, altered early feature access.
Relevance	A criterion-bound estimate of expected downstream control gain if a signal is selected and recurrently re-entered.	Prediction of future switching, calibration, robustness, and policy stability beyond attention alone.
Relevance vs. adjacent constructs	$\mathrm{Rel}_{\mathrm{fb}}$ is not salience, which has no criterion or control-gain requirement; not RPE, which is scalar, one-step, and reward-bound; not precision weight, which modulates prediction error magnitude without recurrent return to latent policy state; not Q-value or advantage, which require a defined policy and reward function rather than criterion-dependent control gain; not empowerment, which measures channel capacity rather than intervention-specific control gain; and not expected free energy, which is not required to be decodable as a separable low-dimensional variable or selectively perturbable.	Each rival predicts $\mathrm{do}(r_t) \approx \mathrm{do}(a_t)$ or $\mathrm{do}(h_t)$ in the four-arm study. $\mathrm{Rel}_{\mathrm{fb}}$ predicts a selectively disproportionate intervention phenotype none of the above require.
Internal feedback	Recurrent re-entry of compressed relevance state into latent processing.	Temporal persistence and downstream modulation of hidden-state trajectories.
Metacognition	Self-evaluative monitoring of uncertainty, confidence, or expected performance.	Confidence reports, opt-out behavior, wager quality, uncertainty heads.
Awareness-like state transition	A change in processing regime produced by recurrent relevance feedback and demonstrable under intervention.	Specific causal effects of perturbing relevance state on control outcomes.
Criterion	The objective or cost structure that defines what counts as important to the system.	Reorganization of latent relevance geometry when criterion changes.

Table 2: Model classes for the artificial experiments. Parameter budgets should be matched as closely as possible.

Class	**Name**	**Key architectural feature**	**Primary test**
M1	Attention-only baseline	Transformer or recurrent controller with standard attention/routing and hidden-state recurrence but no explicit relevance bottleneck.	Whether routing alone is sufficient.
M2	Relevance-augmented model	Baseline plus explicit module that estimates r_t, compresses it, and reinjects it into latent state.	Whether explicit recurrent relevance improves control.
M3	Relevance + metacognition	M2 plus separate metacognitive head or state m_t for confidence/uncertainty estimation.	Whether relevance and confidence dissociate under intervention.
M4	Wider-attention control	Parameter-matched baseline with larger attention capacity but no recurrent relevance bottleneck.	Whether gains come from scale rather than relevance architecture.

Table 3: Primary outcome measures and predicted signatures of relevance perturbation.

Outcome	**Operationalization**	**Predicted effect direction**	**Comparison condition**
Switching cost	Trials or time required to adapt after rule reversal or hazard change.	Relatively larger increase under do(r_t) compared with basic sensory discrimination.	Stronger proportional impairment than matched do(a_t), do(m_t), and equal-magnitude do(h_t).
Calibration error	Gap between confidence and objective success.	Relatively larger degradation under do(r_t) than sensory-discrimination deficit.	Compared against confidence-head or do(m_t) perturbation with proportional spillover quantified.
Distractor susceptibility	Performance loss from salient but task-irrelevant cues.	Disproportionately increased distractor capture under do(r_t).	Must exceed the proportional effect of matched routing perturbation do(a_t).
Policy stability	Variability of action policy under constant latent contingencies.	Relatively larger switching or perseveration under do(r_t).	Not proportionally replicated by equal-variance latent corruption do(h_t).
Long-horizon consistency	Success on tasks in which short-term cues conflict with delayed objective.	Relatively larger impairment under relevance disruption.	Larger proportional impairment than under attention-only perturbation.
Criterion sensitivity	Geometry change in relevance manifold when c changes.	Measurable reorganization for M2/M3 above pre-registered threshold.	Weak or absent in routing-only models and capacity-matched wider-attention controls.
Criterion-invariant subspace stability	RSA correlation of z^* geometry across criterion pairs.	Stable above tolerance threshold.	Full reorganization null model and artificial narrow-criterion baseline.

Note: Predicted effects are relative and proportional, not absolute; some spillover across outcome measures is expected and does not by itself falsify the relevance interpretation.

Table 4: Neural signatures of candidate r_t versus a_t versus m_t.

Feature	**Candidate r_t (relevance)**	**a_t (attention)**	**m_t (metacognition)**
Timescale	Slower decay than a_t predicted only in stable-criterion tasks; difference may shrink or vanish when criteria shift rapidly	Fast, phasic, trial-local	Mixed, often trial by trial
Autocorrelation	Decays more slowly than a_t in stable-criterion tasks	Decorrelates within 1–3 timesteps	Intermediate
Primary regions (decodable)	Prefrontal–insular–striatal circuits as primary candidate substrate; sensory cortex may contain weaker or indirect top-down relevance signals	Parietal, early sensory, frontal eye fields	Medial prefrontal, precuneus, anterior cingulate
Cross-area coherence (candidate)	Secondary correlate that may covary with decoded relevance after movement, arousal, and task-engagement controls	Sensory–parietal–frontal, often gamma-associated	Prefrontal–cingulate
Criterion-shift effect	Measurable geometry change, RSA-detectable above pre-registered threshold	Gain change; less geometry change	Readout recalibration
Perturbation effect (relative)	Larger disproportionate deficit in switching or calibration than in sensory discrimination	Larger disproportionate deficit in early feature selection	Larger disproportionate deficit in confidence or report

Table 5: Proposed experimental program across biological and artificial systems.

Experiment	**Inputs**	**Measurements**	**Expected result / failure mode**
Biological cue-conflict study	Reward cue, threat cue, context cue, interoceptive load, volatility manipulation.	Neural population activity, behavior, physiological state, opt-out or wagering proxy.	Expected: decodable relevance axis predicts switching and calibration; perturbation produces disproportionate control impairment. Failure: axis tracks only arousal or movement.
AI relevance-module study	Identical task family with criterion manipulations under matched parameter budgets.	Performance, calibration, probes, clamping, manifold geometry, OOD robustness.	Expected: relevance-augmented models outperform attention-only baselines on switching and robustness. Failure: benefits disappear after strict scale control.
Cross-domain comparison	Homologous evidence-accumulation and reversal tasks.	Latent geometry, persistence, perturbation susceptibility, criterion classification.	Expected: shared control signatures but domain-specific weighting of threat and urgency. Failure: no common signatures beyond task surface form.
Causal perturbation study	Targeted manipulations of a_t, r_t, m_t, and h_t.	Accuracy, switch cost, calibration, distractor susceptibility, confidence and reaction time.	Expected: proportionally distinguishable intervention phenotypes across arms. Failure: relevance perturbation indistinguishable from generic latent noise or proportional sensory degradation.